DOCUMENT RESUME

ED 060 829 24 JC 720 069

AUTHOR Cohen, Arthur M.; And Others
TITLE Factors Accounting for the Variance in Junior College Students' Composition Writing. Final Report.
INSTITUTION California Univ., Los Angeles. School of Education.
SPONS AGENCY National Center for Educational Research and Development (DHEW/OE), Washington, D.C.
BUREAU NO BR-0-I-051
PUB DATE Jun 71
GRANT OEG-9-70-0028(057)
NOTE 32p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Composition (Literary); *Composition Skills (Literary); English; *English Education; *Junior Colleges; Test Construction; *Testing; *Writing Skills

ABSTRACT

This study explores growth in written composition in the community college by using a group-devised scoring key to score pre- and post-compositions. The study was conducted in three community colleges in Southern California with each student's writing ability being measured by pre- and post-compositions written during the first and last weeks of an 18-week semester. No significant changes in writing ability were detected in this study through a comparison of pre- and post-means for the total sample, or for any of the three colleges, as indicated by a t-test for correlated samples. An analysis of the individual score changes indicated that almost all student scores changed slightly during the semester. In essence, this study supports the use of a cooperatively developed scoring key to reduce rater bias. It does not supportthe assumption that community college students improve their writing skills following 18 weeks of instruction in composition. (Author/MN)

Final Report

Project No. O-I-051
Grant No. OEG-9-70-0028(057)

Factors Accounting for the Variance in Junior College Students' Composition Writing

Arthur M. Cohen
M. Stephen Sheldon
James P. Chadbourne

Graduate School of Education
University of California

Los Angeles, California 90024

June 1971

The research reported herein was performed pursuant to a grant with the Office of Education, U.S. Department of Health, Education and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

U. S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
National Center for Educational Research and Development

Contents

Authors' Abstract

The purpose of the present study was to explore growth in written composition in the community college by using a group-devised scoring key to score pre- and post-compositions. The study was conducted in three community colleges in Southern California, with each student's writing ability being measured by pre- and post-compositions written during the first and last weeks of an eighteen-week semester.

The pre- and post-compositions were collected, given blind code numbers, and randomly selected to form packets of approximately twenty-five compositions. These packets were then distributed for scoring to the twenty-one English instructors in the three colleges. The final sample included 252 pairs of pre- and post-compositions.

Using the scoring key they had developed, the instructors graded their packets of compositions. Group means were computed on a pre- and post-basis for the total sample, as well as for each of the three colleges, on the major section of the scoring key. A cross-tabulation of pre- and post-scores for each of the fifteen items in the scoring key was conducted to analyze changes in writing ability.

The inter-rater reliability of the instructors was checked during the development of the scoring key and again during the final scoring.

No significant changes in writing ability were detected in this study through a comparison of pre- and post-means for the total sample, or for any of the three colleges, as indicated by a t-test for correlated samples. An analysis of the individual score changes indicated that almost all student scores changed slightly during the semester.

This study supports the use of a cooperatively developed scoring key to reduce rater bias. It does not support the assumption that community college students improve their writing skills following eighteen weeks of instruction in composition.

Chapter 1

INTRODUCTION

Does anyone learn to write in college? How would one go about seeking pertinent data? Ask the instructors? Search the dean's files? Poll the students or check their grade point averages?

All available answers suffer from the limitations of bias, distorted perception, and, above all, inadequate information. Grade marks earned in freshman composition courses, for example, say little about writing ability. The classes may be based on traditional grammar, structural linguistics, literature, rhetoric, logic, semantics, communication in the mass media, public speaking, or any combination thereof (Kitzhaber, 1963). In some cases, instruction and practice in writing are not included in the course at all and marks assigned are related instead to verbal performance, responses to quick score exams, or facility in classroom discussion.

Other sources shed no light whatsoever on the question of student learning. Deans' files typically include data on numbers of students who transfer to other institutions and the grades they earn. When students and instructors are queried, their answers are usually, "I feel I learned to write" and "I think our students are writing better now than when they enrolled." The extent to which the questioner is satisfied with those responses depends on his faith in the accuracy of student memory and instructor perception.

If there are answers to these questions, they must be based on certain premises. Among the premises accepted by this study is that colleges are supposed to cause or allow learning to occur. For purposes of this study, learning is defined as changed capability for or tendency toward acting in particular ways. It is assumed that, when students write better compositions at the end of the course than they did at the beginning, they have learned. It is also assumed that the instructor is the person who should assess the level of learning attained by his students. To do this he must define what he will accept as evidence of learning, but individual instructor assessment is not adequate for the measurement of student learning as defined by this study. For a variety of reasons, it is important that the instructors of several classes agree on the way learning shall be measured.

It is not difficult to defend the validity of these assumptions. The definition of learning as "changed capability" is widely accepted (Hilgard and Bower). The pre-post design for assessing learning is also well known (Campbell and Stanley), and the necessity for involving the instructors in determining student learning has been frequently iterated (Scriven). A prime reason for this involvement is

that instruction-wide testing procedures do not often provide information useful in revising the English courses themselves. Many institutions administer entrance examinations and use the results for student placement in various curriculum levels. The student completes a "verbal ability" test or submits a writing sample and is then placed in a course ostensibly geared to his level of proficiency. Any follow-up that may be conducted correlates entrance test scores with the grade marks earned in the courses themselves. This practice, however, has little if any relationship to a student's having learned to write, and fails to yield data useful to the instructors themselves.

Within the schools, composition scoring is typically conducted by individual instructors in the confines of their own courses, but the practice suffers from bias--intentional or otherwise--and frequent distortion. A single instructor's assessment of his students' compositions yields a measure that may be of value to him but that cannot be compared with scores of students in other courses. Each instructor applies his own criteria for his own purposes.

The array of composition scoring devices and procedures currently available makes it difficult to assess student writing ability for purposes of instructional improvement. Many group composition scoring procedures have clear directions and high inter-rater reliability yet yield only global measures (Lambert, 1969). A general rating of the relative worth of a composition may be useful in deciding college admission or curricular placement, but it provides little information on which an instructor may base changes in instructional procedures or emphases.

The problem of determining whether or not (and the extent to which) students learn to write as a result of attending the junior college in particular is far from being resolved. During the past three years, few research documents received and processed at the ERIC Clearinghouse for Junior Colleges address themselves to that issue. The dearth of such studies is noted in several publications resulting from conferences sponsored by the National Council of Teachers of English (Archer, 1965; Braddock, 1963; Weingarten and Kroeger, 1965). These books call for more research on composition scoring, instructional procedures, and variables pertaining to students' learning to write in the junior college. Current pedagogy demands validated procedures for assessing their results (Zoellner, 1969). Research models including designs that can be used by instructors are necessary for these studies.

This design stems from certain definitions of and philosophical positions on education. The over-riding position is that education is a process of moving people from one level of capability or tendency to another. Thus, education is not seen as providing an environment in which something of unknown effect may or may not occur;

rather, education is the bringing about of change. Within the community college, it is the instructors who should predict and define the nature of that change and assess the effects of their instructional process.

A feasible design was prepared and employed by the principal investigator of this project in a study conducted during the 1968-69 academic year. This design was modified for use in this study.

The general hypotheses tested by this research are: (1) feasible procedures for scoring English compositions can be developed; (2) significant differences in writing improvement can be measured and differentiated. The data were collected from sample populations in three colleges: (a) suburban, with 6463 students; (b) urban, with 5432 students; (c) rural, with 1335 students.

Two similar topics were selected for the pre- and post-test. Instructions were simple and uniform. During the first and last week of classes, students received a blue book with instructions to write on the indicated topic. Their compositions were collected and distributed randomly to instructors for rating. Each composition was scored blind and coded on a separate sheet. Pre- and post-compositions were mixed together before being distributed. Scoring sheets provided the basis for analysis. A second reader scored a random 100 compositions from the total sample. An internal consistency of the 16 scores was tested by item analysis. A discrimination co-efficent of .30 was considered satisfactory and .50 was very good.

Chapter 2

PREVIOUS RESEARCH

Overview

It is perhaps misleading to speak of composition research, since the English profession has accumulated far more unanswered questions than it has empirically collected answers. Noting this, Kitzhaber (March 1962, p. 444) has suggested that the English teaching profession urgently needs 15 to 20 years of careful research on the problems of composition. Such a statement is not meant to imply that researchers have completely avoided the problems of composition, but it does indicate the magnitude of the task ahead.

A measure of research activity since 1902 can be obtained by scanning the 504 composition studies cited by Braddock, Lloyd-Jones, and Schoer in their thorough survey, Research in Written Composition. Unfortunately, their review of research uncovered little of value; and the reviewers expressed their disenchantment with the state of the art in this forthright manner:

> Today's research in composition, taken as a whole, may be compared to chemical research as it emerged from the period of alchemy: some terms are being defined usefully, a number of procedures are being refined, but the field as a whole is laced with dreams, prejudices, and makeshift operations (1963, p.5).

It is significant that there has been no shortage of pleas for research from within the English profession (Archer, 1965). The National Council of Teachers of English issued several recent calls for further research on the fundamental questions associated with written composition. Reviews such as Research and the Development of English Programs in The Junior College (Archer and Ferrell, 1965), with its excellent survey of the general problems of teaching English, and English in the Two-Year College (Weingarten and Kroeger, 1965), which spotlights more specifically the difficulties in teaching composition, must certainly be given credit for their searching honesty.

Measuring Change in Writing Ability

Attempts to measure change in writing ability have been generally disappointing; however, in the few studies relevant to this investigation, one or more causes for failure can be identified. Eurich's study in 1932 is a case in point. Using the Van Wagenen English Composition Scale to score 54 freshmen on pre-and post-test themes, he attempted to evaluate the effectiveness of three months of English instruction at the University of Minnesota. His results show that 35 students made no gains, or declined slightly, while 19 made only slight gains. Eurich's conclusion was succinct: "There is no evidence students improve their ability to write in composition" (March 1932, p. 215).

Two points in Eurich's study deserve further comment. First, he notes in his summary of the evidence regarding composition effectiveness that:

> The limited evidence available seems to indicate that improvement may be found primarily on the materials of drill or practice. In other words, general improvement does not necessarily follow specific training. To be effective, training in composition should be directed toward definite and specific ends (Ibid, p. 215).

If the scoring key used in Eurich's study had been employed throughout the period of the experiment, the scoring key might have served to focus student and instructor attention on "definite and specific ends;" with this change in procedure, gains in writing skill might have been identified.

The disappointing results obtained in Eurich's study could also be caused by the short duration of his experiment. It may well be that more than three months of composition instruction is required to cultivate significant changes in writing skill (Scannel and Haugh, June 1968, p.4). Miller (1958) did observe some growth in the writing ability of college freshmen over a period of one year. However, he reports that the improvement after a year of English was no greater than from D+ to C-, and that the majority of his 200 students received the same rating on the first and final papers. It should be noted, however, that the low reliability of the scoring key used to grade pre- and post-themes was a contributing factor to Miller's results.

Related to the findings of Eurich and Miller are the observations of Fellers (1953), who analyzed the type and frequency of composition errors made by 80 students during their last semester in high school against those made by the same students after one semester of college. Fellers found little change in type or frequency of compostion errors, even though all students had taken a college composition course. Thus, while these studies by Eurich, Scannel and Haugh, Miller, and Fellers seem to attest to the persistence of student errors in composition and to the apparent difficulty in causing changes in student writing skills, their conclusions may not be valid when instruction focuses on definite composition skills.

Braddock and Statler (1968) tested the effects of an English writing course versus the effects of no course with 79 matched pairs of freshmen at the University of Iowa over a two-year period. In this study, scoring was done by two groups of raters, each using a different scoring scale; the results showed no significant differences for either group of students. However, these results are equivocal, since neither rating scale produced high rater reliability, and particularly since the raters were able to identify pre- and post-themes. An evaluation procedure must lessen the effects of both these problems.

Other studies of change in writing ability have even reported declines. Sutton and Allen (1964) studied the effect of practice and

evaluation on improvement in written composition. Using control and experimental groups of college freshmen, they elicited twelve themes from each group. Five readers read each theme twice to rate and rank it in relation to the other eleven themes by each writer. Rankings were given on five criteria (ideas, mechanics, wording, form, and flavor) and a five-point scale was used. Sutton and Allen found declines for each group and for the groups combined. The frequency of writing themes that were neither commented on nor returned to the students may have created an attitude of boredom and impatience among the students.

Findings of "no significant difference," as reported by Braddock and Statler and others (1968) need to be examined further, for such a conclusion should not be interpreted to mean that no writing growth occurred. Diederich has commented that the "notation 'not significant' does not prove that there was not true difference between the two things being compared; it only indicates that no true difference was proved" (1964, p. 59).

Some composition studies have made significant breakthroughs in measuring writing ability, if not in fact, at least by suggesting useful techniques. For instance, Dressel, Schmid, and Kincaid (December 1952), while failing to find significant results in their study of the effects of writing frequency, did, however, employ blind scoring procedures that preserved the anonymity of pre- and post-themes. The failure to use a blind scoring procedure appears a major defect in the Braddock and Statler experiment already summarized (1968, p.13). Diederich (April 1966) has also outlined an evaluation procedure that uses scoring blinds.

According to the Diederich plan, the first step toward the improvement of essay grading is to discover how widely instructors within a department disagree when they all grade the same paper without knowing who wrote it or what class level it is supposed to represent. Diederich is not overly optimistic about the attainment of high levels of reader reliability:

> In judging anything as complex as writing ability, however, I think it is unrealistic to expect a higher average agreement in a department than is represented by a correlation of .5 . . . . All that is necessary to get it up to a reliability of .8 is four samples of each student's work, each rated independently by two readers, with a third rating for papers on which there is substantial disagreement (April 1967, p. 582).

Nevertheless, through a process of evaluation and discussion among raters, Diederich suggests that a department could achieve "reasonably uniform standards in grading" in about three years (Ibid, p.584). He also asks a provocative question that highlights the need for efforts to achieve consensus on grading criteria: "I wonder why we should pretend to be able to teach anything like good writing if no two of us can agree even this much on what it is" (Ibid, p. 585).

Kincaid (1953) studied factors affecting variations in the quality of student writing and found that neither the content factor, because of different assigned topics, nor the pressure introduced by the examination situation had any significant effect on the average quality of writing by student groups of twenty or more. He also found that a single paper could provide a valid basis for evaluating writing ability and that more reliable information could be obtained from a single pre-test theme and a single post-test theme for measuring the overall, or average, group improvement.

Diederich (December 1944) attended to the problem of essay topics. He suggests that student writing can be measured provided:

1. topics assigned are within the student's comprehension, but not so easy that levels of excellence cannot be determined

2. the form of writing assigned represents the kind of writing students may be expected to use in later life

3. all students write on the same topic, and all papers are based on a common set of materials

4. the essay is written in class

5. if the mark on the exam is of vital importance to the student, the composition is read by two readers

6. the papers are read without the reader's knowing who wrote them

7. the papers are marked in accordance with criteria formulated and written down in advance

8. for individual measurement, at least two essays on different topics are written to give a reliable measure of skill in writing.

Buxton's study (1958) supports the suggestion made by Diederich and others that rater reliability can be increased by experience in the development of and practice with a scoring key. Buxton also supports the use of class and student blinds as a means of insuring rater reliability in essay grading; unfortunately, his results cannot be relied on, for his raters apparently were aware of which were pre- and which were post-essays.

The Problem of Scoring Essays

In all the attempts to measure change in writing ability noted here, two problems recur: the problems of reader reliability and establishing scoring criteria. Reader variability in the scoring of essays is well documented and may be regarded as the chief obstacle

to the use of written compositions for measuring change in writing ability. Englehart (1969, p. 407) notes that the unreliability of readers has been recognized since the 1880s. Since that time, numerous other researchers have encountered the same problem (Starch and Elliott, 1912; Fostvedt, November 1965; Jewell, Cowley, and Rhum, 1966), yet, despite wide recognition of the problem, few successful solutions have been advanced. Even one of the most experienced composition researchers, Braddock, has stated that most people know that the grading of compositions is notoriously unreliable. While there are undeniable difficulties in securing reader reliability, the accuracy of Braddock's statement deserves further inquiry.

The efforts of Diederich, French, and Carlton (1961) well illustrate the phenomenon of reader variability. Using three hundred college themes, they asked 53 judges in the fields of English, social studies, natural sciences, law, writing and editing, and business to judge each theme on a nine-point scale. The results show that "ninety-four percent of the themes received either seven, eight, or nine of the nine possible grades and no paper received less than five different grades from the 53 readers: (Ibid, p.58).

While these results show a wide range of reader variability, it is significant to note that the ten English instructors in the study had a higher mean intercorrelation (.41) with one another than did any other of the occupational groups. One might well ask of what use is it to know that lawyers and business executives cannot agree on the excellence of student compositions. It is significant, however, to note that none of the readers was given criteria by which to rate the compositions in this study. Very different results might have occurred if standard criteria for scoring had been employed.

Other researchers have supported the use of essays for measuring writing skill, yet their solutions to the problem of reader variability have often lacked practicality. Greene and Petty (1963) have suggested that reliable essay evaluation requires extensive samples as well as repeated ratings by expert judges to minimize variability. Godshalk, Swineford, and Coffman (1966) reached similar conclusions, for they report that essay-score reliability is a function of the number of different essays and the number of different readings. They recommend the evaluation of five different essay topics, with each topic read by five different readers. It is obvious that such an evaluation procedure would take too much time; it is therefore not a practical solution to the problem of reader variability in the community college.

The work of Follman and Anderson (1967) offers provocative suggestions to ameliorate the unreliability of scoring keys. Using upper-division college English majors as raters, Follman and Anderson assigned ten themes to five groups of raters to determine the intra-reliability of five composition scales. Follman and Anderson found that the differences among rating groups did not change with the subject matter of the essays, and that the essays received substantially the same scores from all five rating groups. Their conclusion is that the high reliability across different evaluation procedures "may be due primarily to the homogeneous nature of the raters rather than to a rating system" (Ibid, p.199)

This conclusion is particularly notable since one of the evaluation procedures, the Everyman's Scale, allowed each rater to use his own criteria--yet the second highest reliability score was obtained using the Everyman Scale. Follman and Anderson observe that:

> It may now be suggested that the unreliability usually obtained in the evaluation of essays occurs primarily because raters are to a considerable degree heterogeneous in academic background and have different experimental backgrounds . . . likely to produce different attitudes and values which operate significantly in their evaluations of essays (Ibid, pp. 198-199).

Their comment on the value of scoring keys is significant:

> The function of a theme evaluation procedure, then, becomes that of a sensitizer or organizer of the rater's perception and gives direction to his attitudes and values; in other words, it points out what he should look for and guides his judgment (Ibid, p.199).

This suggests the possible value of cooperative grading practices by English departments in the community college. Indeed, Diederich has written about the possible values of cooperative grading of English compositions:

> It makes the job easier, quicker, and more interesting by a division of labor; it puts teachers and students on the same side of the fence; it reveals answers to many teaching problems; it provides ammunition against our critics; and it adds fun and excitement to both teaching and learning (1967 , p.579).

Nealy (November 1969) even demonstrated that standards for composition are "readily communicable" between instructors and their students. Surely English instructors within a department ought to be able to increase their reliability through practice with standard scoring criteria, despite the initial difficulties of reaching consensus.

The potential merits of composition scales bear repeating in 1970:

> These scales are but means toward ends. The most important of these ends are: (1) to test impartially the various methods of teaching composition by measuring their results; (2) to measure these results in accurate, objective, stable, and understandable terms; (3) to furnish a common basis for comparing the same class or school or that of pupils in different classes or schools; (4) to classify pupils fairly in compostion; (5) to grade them justly within their group; (6) to enable teachers to discover their reliability in judging the general merit of English Composition; and (7) to furnish pupils an incentive to self-competition (Hudelson, 1925).

The literature on composition scales is clearly not definitive. The development of practical scoring criteria is needed, for, as Diederich comments:

> I honestly believe that almost all experiments concerning English compostion that rely on essay grades have been conducted with tape measures printed on elastic, and that they must be replicated with measures in which we can have confidence (1964, p.60).

The question remains: What type of scale is superior? Studies such as those by Cast (1939 , 1940), Coward (1952), and Nisbet (1955) have attempted to determine the superiority of the analytic versus the holistic approach to grading essays. While Nisbet found that the two methods could produce nearly equivalent results in terms of rater reliability, both Cast and Coward found some evidence of the superiority of the analytic approach. The scoring method selected should depend on institutional purpose and philosophy: in the community college, with its diverse student population, there is a detailed information about program evaluation and improvement. The holistic approach to grading essays simply will not yield as much useful information as the atomistic approach.

Other researchers have investigated different point scales. McColly and Remstad (October 1965) demonstrated that a four-point scale can be used with reliability equal to that of the more time-consuming six-point scale. Jewell's (1966, p.19) work supports the utility of the four-point scale for scoring essays.

The use of explicit scoring criteria seems to offer the most promise for reducing reader variability in the scoring of written compositions. Studies by Torgerson and Green (1953), Diederich, French, and Carlton (1961), Nybert (1966), and Hyndman (1969) have explored the identification of clusters of grading variables. Their studies should have led to further experiments with scoring keys. The fact that few recent studies have employed scoring keys may be attributed to the persistent belief in their unreliability. It may also be true that the persistent belief in the unreliability of scoring keys stems from a widespread failure to experiment with composition scales.

Summary

This chapter has reviewed composition research related to the problem of measuring changes in writing ability. Both difficulties and potential solutions have been noted. The major problem confronting the composition researcher is to reduce reader variability. The use of blind scoring procedures, of a four-point scale, of a cooperatively devised scoring key with clearly defined criteria, of pre- and post-themes, the use of practice in grading, and the use of relatively inoffensive experimental procedures have been identified as potential aids in securing meaningful measures of writing growth. In Chapter 4, attention will be given to each of these concepts.

CHAPTER 3

RATIONALE AND DESIGN

Rationale

This design stems from certain definitions and philosophical opinions regarding education. The overriding position is that education is a process of moving people from one set of capabilities or tendencies to another. Within this framework are certain definitions; e.g., learning is changed capability for, or tendency toward, acting in particular ways; and instruction is the deliberate sequencing of events so that learning occurs. Thus, education is not seen as providing an environment in which something of unknown effect may or may not occur; rather, it is the bringing about of change. Within the community college, it is the instructors who should predict and define the nature of that change and assess the effects of their instructional process, yet, despite the notation in every college catalog, "The student will learn to write effectively," no one really knows the extent to which writing improves, or if, in fact, it improves at all as a result of college attendance.

Specifically stated, nine major assumptions underlie this study:

1. Community college English instructors have an obligation to define and assess the effects of their instruction.

2. Learning is shown by changed capabilities.

3. It is possible to measure reliably changes in writing ability through a comparison of pre- and post-themes written a semester apart.

4. Community college English instructors can devise a valid scoring key; with practice, they can use the scoring key to grade English compositions reliably.

5. Such a scoring key should contain several scoring categories.

6. The scoring key should call for analytic distinctions, rather than global judgments.

7. Rater reliability should be measured through an analysis of variance, using scores assigned to common compositions.

8. Scoring blinds (student, college, class, time) assist in achieving objective scoring of student themes.

9. The random assignment of papers from different classes to be scored by many scorers helps to distribute the influence of any single reader.

The design of this study is an attempt to know more about the

effects of composition courses and to solve the major problems associated with the measurement of writing skills. First, the design calls for actual demonstrations of composition skill, and, in this respect, has greater face validity than studies that rely on indirect measures of writing skill. The very fact of pre- and post-testing allows a defensible comparison of compositions and permits the measurement of change in writing ability.

Second, the design mitigates the causes of reader unreliability in the scoring of essays. Through the use of scoring blinds, reader bias regarding authorship, class level, and time of composition are controlled.

Third, error introduced by difficult or easy readers is controlled by random sampling in the distribution of themes for final scoring. In this way, scoring error between the pre- and post-tests is balanced.

Fourth and finally, variations in scoring are mitigated through the definition of and practice with explicit scoring criteria. The design reflects the belief that a meaningful assessment of composition skill requires the cooperative evaluation of essays according to clearly defined criteria (Diederich, 1967). Through the use of the cooperatively developed scoring key, reliable measures of writing ability are obtained and, since these measures are guided by a common scoring key, it is possible to make comparisons between the writing performnances of students in English composition programs in different colleges.

Hypotheses

Two general hypotheses were tested by this research. Each implies a series of specific hypotheses:

1. Procedures for scoring English compositions can be developed that are reliable, internally consistent, have face validity, and are meaningful for the evaluation and change of instruction.

2. There are significant differences in the improvement in writing ability between students enrolled in different colleges.

Design Procedures

To seek an assessment of writing ability, student compositions were written during the class periods in the first and last weeks of the fall semester at three community colleges. Pre- and post-compositions were collected and matched and given blind code numbers. Bluebook covers that identified the student and college were removed. Pre- and post-compositions were mixed, and packets of approximately twenty-five compositions were then randomly formed and distributed for scoring.

Score sheets were collected by the investigator, and the data were then key-punched for processing by computer to determine group means on the pre- and post-compositions. A series of Chi-squares were computed to measure individual and group scoring changes. Pearson correlation co-efficients were computed for the scoring key variables. A test of rater reliability was performed on the basis of three common themes that were placed in each instructor's packet of themes.

Selecting the Participants

The presidents and deans of instruction of several Southern California community colleges were asked to participate in this study. Three colleges expressed an interest and the design was presented to interested faculty members at each college. The only restriction placed on faculty by the investigator was that they be assigned to teach a freshman writing course.

Because one college had considerably more participating instructors than the other two, most of the themes came from that institution.

Selecting the Classes

Participating instructors were asked to select one or more of their classes for use in this experiment, depending on the number of assigned writing courses they were scheduled to teach during the fall semester, 1969-70. Most of the student themes in this study came from the standard English course for freshman students.

Developing the Scoring Key

Building on the work of previous research, instructors from only one college constructed the scoring key, which calls for ratings on each of fifteen factors forming three general categories: content, organization, and mechanics. (See Scoring Key, Figure I) The four-point scale was thought to assist instructors to score themes faster than and just as reliably as a scale calling for finer distinctions. The scoring ranges from zero for a perfect score to three for unacceptable performance for each of the fifteen scoring items.

Checking Rater Reliability

During the development of the final scoring key, reader reliability was checked by having each instructor read and score duplicated compositions. Scoring items that failed to produce reader agreement more than 75 per cent of the time were modified or rejected. After three reliability trials and before the final scoring, all scoring items met this requirement. During the final scoring, each instructor and the investigator rated three duplicated themes as a final measurement of reader reliability.

Choosing the Topic

The two topics selected for the pre- and post-themes were: "What makes a good advertisement?" and "What makes a good entertainer?" These

topics were selected for several reasons. They did not ask for responses that might be intensely personal; they did not invite trite responses; they did not require the student to understand rhetorical terms such as "compare and contrast"; they both called for expository writing.

Instructions to Students

Clear instructions to students were printed on sheets of paper pasted on the covers of the composition books. For each of the two composition topics, the instructions were:

> You are to write a composition in this blue book.
> Write in ink on one side of the paper only.
> Write on alternate lines, please.
> Your topic is: What makes a good entertainer?
>
> [Repeat the same instructions.]
> Your topic is: What makes a good advertisement?

Administering the Composition

During the first week of his course, each instructor randomly distributed to his class or classes a number of blue books; approximately half of each class then wrote for a class period on each of the topics for the pre-test themes.

During the last week of the course, the instructor distributed to his classes blue books on which the investigator had written student names from the pre-tests, insuring that each student wrote his post-test composition on the topic different from his pre-test.

Coding, Sampling, and Scoring

All pre- and post-themes were collected and identifying marks from each composition were removed. Names were entered on a master sheet and a code number was stamped on the blue book. Code numbers were assigned on a random basis to prohibit number sequence from identifying which themes were pre- and which were post-test.

Approximately 500 themes were matched on a pre- and post-basis because of student attrition. To reduce the correction load for each instructor, 252 pairs of themes were randomly selected for final scoring. The results of this study apply only to those students who persevered throughout the entire semester of instruction.

Packets of approximately 25 themes, plus three duplicated themes for the reliability check, were then distributed for scoring to each participating instructor. Each instructor was asked to score his packet of compositions according to the criteria in the cooperatively developed scoring key.

Analysis, findings, and results of this study are discussed in Chapter 4.

CHAPTER 4

RESULTS

A. Measurement of Rater Reliability

While this investigation proposed to measure changes in the writing ability of a sample of community college students, it also examined the measurement of reader reliability. When error variance or rater variance accounts for more of the total variability than does the quality of the essays, little can be said about the growth in writing ability of students. Reliability in this experiment varies according to the accuracy or sameness with which various instructors use the scoring key.

Three measures of reliability were used for this experiment. One was an approximation of a reliability coefficient from the standard error of measurement formula; a second was a correlation between an independent rater with all other raters on 100 randomly selected essays; and the third was an analysis of a comparison of the variance due to the rater and error with the variance due to the quality of the essays.

As part of the procedure, every participating English teacher (i.e., rater) rated three common essays. The variability in their total score for these essays as measured by standard deviation was 7.46, 7.79, and 7.93. The average was 7.8 One can reasonably assume this average standard deviation of 7.8 to be a good estimate of the standard error of measurement for this scale. As a consequence, a reliability coefficient can be obtained by working backward using the formula.

$$S_e = S\sqrt{1-r}$$

where

S_e = standard error of measurement

S = standard deviation of the test

r = reliability of the test

Through computation with the standard error of measurement on the scores given to the three common themes and the standard deviation for total pre- and post-test scores, we can obtain an estimate of reader reliability:

$$S_e = S\sqrt{1-r}$$

$$7.8 = 10.73\sqrt{1-r}$$

$$\sqrt{1-r} = .727$$

$$1-r = .529$$

$$r = .471$$

A second estimate of reliability coefficient was obtained by collecting two readings of a random sample of 100 essays. For this reliability study, 100 essays were randomly drawn from the 504 essays used in the study. These represent all three colleges, both pre- and post-test. Each of the raters had read and scored some portion of the sample. They were re-read and independently scored by an additional rater. The correlations between this independent rater and the first reading are given in Table 1.

Table 1

Correlation: One Reader With All Others

N = 100

Content	.39
Organization	.41
Mechanics	.32
Total	.36

Essays randomly drawn from total population of essays, both pre- and post-test.

The third procedure for estimating reliability, though unorthodox, has most meaning for this type of research. It results in separating the overall variance into that portion due to the quality of the essays and that portion due to instructor bias plus random error. Using the scores assigned by each instructor to his sample of essays, means and standard deviations were computed for each who has read 23 or more essays. There were 18 such instructors. This was done for content, organization, mechanics, and total. The variance of these instructor means was then computed for each sub-area and the total. The results appear in Table 2. Remembering that each sample of essays given to each instructor was randomly selected, the variance of these means gives a good estimate of the proportion of the total variance due to instructor bias plus sampling error. When this variance is subtracted from the total variance of the 504 essays, one can determine with fair accuracy how much was attributable to the quality of the essays. The proportion of variance due to instructor bias and error can be determined by dividing the variance of the instructor means by the total variance. These data are also reported in Table 2.

B. Measurement of Group Progress

The results of the comparisons between total sample means pre- to post-, are shown in Table 3.

TABLE 2

RESULTS OF PARTIALING OUT VARIANCE DUE TO ESSAYS AND VARIANCE DUE TO INSTRUCTOR PLUS ERROR.

	$\bar{X}$	S	S^2	S^2 Inst.	% S^2 total due to Inst.	% S^2 total due to Essays
Content	4.68	2.37	5.62	1.04	19%	81%
Organization	11.23	5.53	30.58	6.03	20%	80%
Mechanics	6.53	3.96	15.68	2.32	15%	85%
Total	22.50	10.73	115.13	13.65	12%	88%
	N = 504			N = 18		

DATA FOR COMPUTING INSTRUCTOR VARIANCE

N = 18

$\bar{X}_1$	S_1	$\bar{X}_2$	S_2	$\bar{X}_3$	S_3	$\bar{X}_4$	S_4	Instr.
1.44	2.04	3.04	4.03	2.56	3.11	7.04	8.52	01 N=25
5.76	1.62	13.30	3.47	6.64	2.36	25.70	6.15	02 N=33
4.35	2.78	10.62	6.13	7.62	4.65	22.59	11.85	03 N=29
5.30	2.02	14.07	5.17	6.83	3.44	26.20	9.75	04 N=30
5.42	2.62	12.33	5.90	6.67	3.47	24.42	10.77	05 N=24
3.63	1.93	8.83	3.86	3.96	3.09	16.42	8.26	06 N=24
4.48	1.81	9.41	4.84	3.97	3.80	17.86	9.51	07 N=29
4.54	1.99	10.73	4.31	6.81	3.26	22.08	8.50	09 N=26
4.88	2.73	10.72	5.49	7.12	5.08	22.72	11.51	10 N=25
5.03	2.15	11.84	4.64	8.09	3.36	24.97	9.89	11 N=32
5.33	2.23	13.80	5.33	6.63	4.56	25.77	10.89	12 N=30
4.96	2.52	15.08	5.66	8.00	4.53	28.04	11.80	13 N=26
3.89	1.21	9.04	2.99	6.39	3.10	19.31	5.94	14 N=26
3.77	2.70	10.73	6.19	7.96	4.52	22.46	12.97	16 N=26
4.83	1.97	10.25	5.09	6.21	3.23	21.29	8.18	17 N=24
4.42	1.77	10.53	3.59	5.50	2.53	20.42	7.10	18 N=26
5.37	2.55	11.83	6.16	8.17	4.54	25.37	12.51	21 N=30
6.00	2.33	15.26	4.85	6.83	3.33	28.09	9.35	23 N=23
N 18		18		18		18		
$\bar{X}$ 4.63		11.19		6.44		22.26		
S^2 1.04		6.03		2.32		13.65		

TABLE 3

MEANS, PRE-TO POST-

N = 252	PRE-		POST-		
	Mean	SD	Mean	SD	Dif.
Content	4.8	2.4	4.6	2.4	.2+
Organization	11.2	5.5	11.3	5.6	.1-
Mechanics	6.8	4.0	6.2	3.9	.4+
Total	22.8	10.8	22.1	10.7	.73+

Since a score of zero on the rating scale indicates a perfect score, a zero mean would indicate a perfect mean score. Thus, declines in mean scores from pre-to post-test indicate gains in writing ability. A comparison of total sample means, pre- to post-, indicates that a slight but insignificant gain in writing ability was measured (.73+).

An inspection of the pre-to post- changes in total sample means by major subsections of the scoring key shows that slight gains were observed for Content (.2+) and for Mechanics (.4+), while Organization shows a slight decline (.1-). The magnitude of these mean changes is not significant enough to indicate that any true gains in writing ability were detected in this experiment.

A comparison of pre- and post-test means for each of the individual colleges reveals no significant differences. Table 4, 5, and 6 report these data. Table 4 shows a comparison of pre- and post- means for College A on the major subsections of the scoring key.

TABLE 4

COLLEGE A: MEANS, PRE- TO POST-

N = 191	PRE-		POST-		
	Mean	SD	Mean	SD	Dif.
Content	4.7	2.4	4.6	2.4	.1+
Organization	11.1	5.5	11.4	5.6	.3-
Mechanics	6.5	4.0	5.8	3.8	.7+
Total	22.3	10.8	21.8	10.7	,5+

Table 5 shows a comparison of pre- and post- means for College B on the major subsections of the scoring key.

TABLE 5

COLLEGE B: MEANS, PRE- TO POST-

N = 24	PRE-		POST-		
	Mean	SD	Mean	SD	Dif.
Content	5.5	2.0	5.5	2.4	.00
Organization	12.2	4.3	13.1	4.8	.9-
Mechanics	9.0	3.7	8.9	4.2	.1+
Total	26.8	8.0	27.5	10.0	.7-

Table 6 shows the comparison of pre- and post- means for College C on the major subsections of the scoring key.

TABLE 6

COLLEGE C: MEANS, PRE- TO POST-

N = 37	PRE-		POST-		
	Mean	SD	Mean	SD	Dif.
Content	4.6	2.4	4.1	2.1	.5+
Organization	11.1	6.3	9.7	5.8	1.4+
Mechanics	7.2	4.1	6.6	3.5	.6+
Total	22.9	11.8	20.4	10.4	2.5+

While the preceding results show insignificant changes in group means, pre- to post-, an analysis of changes item by item on the scoring key, pre- to post-, for the total sample reveals considerable movement. Table 7 presents these data. An inspection of Table 7 shows that only one-third of the students in the sample showed no change from their pre- to post-composition scores. The data from this analysis of change suggest that the previous comparisons of group means, pre- to post-, mask the full extent of the

changes that took place in student scores. In the Content subsection, 4% more students showed gains than declines according to changes in totals for this section. In the Organization subsection of the scoring key, a comparison of totals shows that an equal number of students gained and declined; only 3% of the sample showed no change in scores from pre- to post-test in this section. The Mechanics subsection shows that only 1% more of the sample declined than gained; 11% of the students made no changes in their scores on this section from pre- to post-test.

TABLE 7

CROSS-TABULATION OF SCORING ITEMS,

PRE-TO POST-, FOR THE TOTAL SAMPLE

N = 252

Scoring Items	% Gain	% Decline	% No Change
1 Treatment	38	33	29
2 Knowledge	37	31	32
3 Diction	35	32	34
CONTENT TOTALS	45	41	14
4 Thesis	39	36	25
5 Plan	33	38	29
6 Unity	25	44	31
7 Development	33	34	33
8 Patterns	32	32	36
9 Transitions	32	32	36
10 Logic	36	34	30
ORGANIZATION	47	47	3
11 Spelling	40	27	33
12 Syntax	37	28	35
13 Punctuation	36	35	29
14 Major error	38	30	32
15 Minor error	36	31	33
MECHANICS TOTALS	44	45	11
GRAND TOTALS	49	48	3

Pearson correlation coefficients were computed for scoring-key items to measure the validity of each item in the scoring key. In addition, Pearson correlation coefficients were computed for scoring-key items to total test scores to provide measures of the rater's practice in assigning scores. Table 8 shows the degree to which scoring-key items were related to total scores on the pre- and post-test. Only three scoring items show less than a .72 relationship to the

TABLE 8

CORRELATION COEFFICIENTS FOR SCORING ITEMS TO TOTAL TEST SCORES, PRE- TO POST-

	Scoring Item	Total Score Pre-test	Total Score Post-test
1	Treatment	.7838	.7709
2	Knowledge	.7691	.7827
3	Diction	.8101	.8485
4	Thesis	.7104	.7359
5	Plan	.8240	.8377
6	Unity	.8150	.8497
7	Development	.7361	.7598
8	Patterns	.8455	.8335
9	Transitions	.8290	.8348
10	Logic	.7946	.8241
11	Spelling	.6098	.5345
12	Syntax	.8021	.8031
13	Punctuation	.7360	.7882
14	Major error	.7275	.6723
15	Minor error	.6916	.6417

total scores: Item 11 Spelling, Item 14 Major Mechanical Errors, post-test only, and Item 15 Minor Mechanical Errors. The complete scoring key and criteria appear in Figure I.

Tables 9 and 10 show the intercorrelations between scoring-key items, on the pre- and on the post-tests, an analysis useful in identifying which items may be of little use in the scoring key.

Chapter 5 presents the investigators' conclusions and suggestions for further study.

TABLE 9

INTERCORRELATIONS OF SCORING-KEY ITEMS, PRE-TEST

	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	.87	.61	.57	.73	.69	.64	.59	.61	.61	.36	.50	.45	.42	.40
2		.61	.56	.71	.69	.63	.58	.57	.59	.34	.49	.47	.38	.41
3			.54	.61	.57	.57	.69	.65	.63	.48	.73	.56	.56	.53
4				.70	.68	.52	.49	.56	.54	.36	.48	.41	.36	.37
5					.85	.68	.62	.68	.66	.37	.54	.46	.46	.42
6						.69	.62	.69	.66	.39	.54	.42	.45	.41
7							.65	.61	.55	.32	.51	.43	.36	.33
8								.76	.65	.49	.70	.66	.64	.59
9									.75	.45	.63	.53	.55	.52
10										.38	.64	.53	.51	.44
11											.53	.47	.52	.49
12												.65	.65	.61
13													.70	.65
14														.73

TABLE 10

INTERCORRELATIONS OF SCORING-KEY ITEMS, POST-TEST

	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	.81	.66	.63	.72	.72	.67	.57	.61	.63	.30	.50	.49	.32	.31
2		.69	.62	.69	.72	.68	.59	.62	.66	.28	.52	.48	.41	.29
3			.59	.72	.70	.58	.67	.65	.70	.40	.75	.65	.53	.52
4				.70	.69	.58	.56	.59	.59	.27	.46	.48	.37	.32
5					.88	.67	.64	.68	.70	.34	.56	.54	.49	.38
6						.69	.67	.69	.71	.37	.58	.56	.43	.36
7							.62	.64	.62	.30	.52	.51	.37	.30
8								.75	.65	.44	.67	.68	.54	.56
9									.76	.35	.63	.62	.51	.53
10										.30	.68	.62	.51	.40
11											.46	.42	.37	.49
12												.74	.60	.61
13													.65	.62
14														.65

CHAPTER 5

CONCLUSIONS

Significant changes in writing ability were not detected in this study through a comparison of pre- and post-means for the total sample, or for any of the three colleges. A t-test was used to determine the significance of the differences between the means for the correlated samples in this study.

An analysis of the individual score changes, pre- to post-, on the fifteen scoring items revealed that nearly all student scores changed during the semester.

This experiment supports the use of a scoring key to enable English instructors to achieve a high degree of consistency in the grading of essays. Most of the observed differences in student compositions are apparently true differences and are not attributable to inter-rater unreliabilty. Further, an analysis of the correlations between the scoring items and total test score indicates the homogeneous nature of instructor scoring patterns. The relationships between item scores and total test scores on the pre-test approximate those ahieved on the post-test, with one exception: Item 11 Spelling Errors.

Interpretation of Findings

It is possible that the four-part scale contributed to the finding of no change in writing ability in this study. It is also possible that the four-way scale allowed for too much discrimination, or too wide a range of instructor choice. Another possible explanation for the results is that, since most of the students were in the standard English course, there were no sharp differences in skill level.

The most likely explanation is that students do not learn to write during eighteen-week courses that concentrate on numerous English-language activities. Extensive practice in expository writing, with topics similar to those used in this study, might well cause significant changes in writing ability if this were the foremost goal of the course. The amount and type of writing called for by each of the mixed-bag courses in this study vary. It remains to be tested if a semester of instruction, concentrated on the problems of expository writing, can cause significant gains in writing ability.

It is also possible that the results can be attributed to the one-shot nature of the experiment. It is not known from this study what the results would be if English instructors and their students worked with a similar scoring key throughout a semester or for longer periods of time. Student knowledge of the scoring-key items might well assist them to develop their own writing ability. The problem of student motivation also remains open to question. While students and instructors indicated enthusiasm for the project, they may have regarded the

two essays in this project as peripheral to the instructional program in their classes. Still the possibility that students do not learn to improve their writing skills must be admitted.

In any event, further experiments with this scoring key and these procedures are needed to establish a baseline of possible changes that might be attributed to highly effective instruction. Even if significant changes had been found through a comparison of group means, it would not be possible to call the degree of change satisfactory or unsatisfactory without replication.

A significant finding is that a high degree of rater reliability can be achieved through the use of a scoring key. It is possible that the cooperative development of the scoring key and the subsequent practice in its use contributed to the high degree of reliability attained. The reliability of the readers' scores was also increased by the scoring blinds built into the evaluation procedures. These efforts to limit reader bias appear to have been highly successful.

Rater reliability was also raised by the random sampling procedures used to distribute pre- and post-tests. Whatever scoring variation existed between the raters on the pre-tests, it can be assumed that it also existed during the post-test scoring; in effect, high scorers and low scorers cancelled each other out through the random distribution of the essays for grading. While this procedure undoubtedly increased the objectivity of the results, it also may have contributed to the leveling of the gain scores. The utility of broad scoring categories is not supported by this investigation, since no significant gains in writing ability were determined on the basis of total scores or on subsection scores.

In summary, the high degree of inter-rater reliability estimated in this study and the results obtained support the conclusion that the students did not significantly improve their writing abilities during the eighteen-week period of instruction. While a variety of factors may have contributed to these results, the evidence calls into question studies in which multiple blind-scoring techniques were lacking and in which substantial gains in ability were recorded. Rater bias simply cannot be controlled when readers know whether the compositions were written before or after instruction.

Strenghts and Weaknesses of the Study

A major strength of this study is that it demonstrates that

reader reliability can be improved through the use of a cooperatively developed scoring key. The study also supports the use of these evaluation procedures as an in-service training program. While their benefits in in-service training are not directly observable in this study, the fact that twenty-one English instructors from three colleges could attain high reliability suggests the value of cooperative grading of essays.

Another strength of this design is the relative ease of carrying out the procedures once a scoring key is developed. The amount of time required to collect, identify, and redistribute the compositions for scoring is minimal and need not cut into instructor time for these duties can be delegated to clerical help. The data analysis can be programmed easily by the director for research and his staff. It is possible, excluding the development of the scoring key, that the actual scoring of themes would take less than the normal time, since the instructor would know precisely what he is looking for in the compositions. In any event, the use of this evaluation technique as an in-service training device and as a way to evaluate instruction is more defensible than current grading practices, which yield few benefits to anyone.

This study has two important weaknesses. Because instructors handled a variety of writing problems during the semester, it cannot be stated that instruction concentrated on the type of writing evaluated in this study. Because the study was coordinated by outsiders, this weakness could not be lessened. Even under the best of conditions, it might be difficult to find an English Department that would agree to emphasize expository writing almost exclusively. If this had been done, however, the results of the study might be quite different.

The second weakness is that the problem of student motivation was not given enough attention. While the test topics were relatively easy, no attempt was made to motivate the students to write their best for the study. Only in College C did the instructors treat the writing as a graded part of their regular classwork. If the post-tests had been the culminating exercise in the course at each of the colleges, the results of this study might again have been quite different.

REFERENCES

Archer, Jerome W. anf Ferrell, Wilfred A. Research and the Development of English Programs in the Junior College. Champaign, Illinois: National Council of Teachers of English, 1965.

Braddock, Richard, Lloyd-Jones, Richard, and Schoer, Lowell. Research in Written Composition. Champaign, Illinois: National Council of Teachers of English, 1963.

Braddock, Richard and Statler, Charles R. Evaluation of College-Level Instruction in Freshman Composition, Part II. United States Office of Education Cooperative Research Project No. S-260. Iowa City: University of Iowa, 1968.

Buxton, Earl W. "An Experiment to Test the Effects of Writing Frequency and Guided Practice Upon Students' Skill in Written Expression." Unpublished doctoral dissertation, Stanford University, 1958.

Campbell, Donald T. and Stanley, Julian C. "Experimental and Quasi-Experimental Designs for Research on Teaching." In Handbook of Research on Teaching, edited by N. L. Gage. Chicago, Rand McNally, 1963.

Cast, B. M. D. "The Efficiency of Different Methods of Marking English Composition," British Journal of Educational Psychology, 9 (November 1939), 257-269; (November 1940), 257-269.

Coward, Ann F. "A Comparison of Two Methods of Grading English Compositions," Journal of Educational Research, 46 (1952), 81-93.

Diederich, Paul B. "Cooperative Preparation and Rating of Essay Tests," English Journal, 56 (April 1967), 573-590.

Diederich, Paul B. "How to Measure Growth in Writing Ability," English Journal, 55 (April 1966), 435-449.

Diederich, Paul B. "The Measurement of Skill in Writing," School Review, 54 (December 1944), 584-592.

Diederich, Paul B. "Problems and Possiblities of Research in the Teaching of Written Composition," Research Design and the Teaching of English, David H. Russell, director. Champaign, Illinois: National Council of Teachers of English, 1964.

Diederich, Paul B., French, John W., and Carlton, Sydell T. Factors in Judgments of Writing Ability. Research Bulletin RB 61-15. Princeton, New Jersey: Educational Testing Service, 1961.

Dressel, Paul B., Schmid, John, and Kincaid, Gerald. "The Effect of Writing Frequency Upon Essay-Type Writing Proficiency at the College Level," Journal of Educational Research, 46 (December 1952), 283-293.

Englehart, Max D. "Examinations." In Encyclopedia of Educational Research, edited by Robert L. Ebel. 4th edition. New York: Macmillan, 1969.

Eurich, Alvin. "Should Freshman Composition Be Abolished?" English Journal, 21 (March 1932), 211-219.

Fellers, Alvin L. "Problems in Writing in College Composition Classes." Unpublished doctoral dissertation, Stanford University, 1953.

Follman, John C. and Anderson, James A. An Investigation of the Reliability of Five Procedures For Grading English Themes. Champaign, Illinois: National Council of Teachers of English, 1967.

Fostvedt, Donald R. "Criteria for the Evaluation of High School English Composition," Journal of Educational Research, 59, No. 3 (November 1965), 108-112.

Godshalk, Fred I.; Swineford, Francis; and Coffman, William. The Measurement of Writing Ability. New York: College Examination Board, 1966.

Greene, Harry A. and Petty, Walter T. Developing Language Skills in the Elementary Schools. Boston: Allyn and Bacon, Inc., 1963.

Hilgard, Ernest R. and Bower, Gordon H. Theories of Learning (3rd ed.), New York: Appleton Century Crofts, 1966.

Hudelson, Earl. "The Effect of Objective Standards Upon Composition Teachers' Judgments," Journal of Educational Research, 12 (1925), 329-340.

Hyndman, Roger. "Some Factors Related to the Writing Performance of Tenth-Grade Students." Unpublished doctoral dissertation, University of California, Los Angeles, 1969.

Jewell, Ross M., Cowley, John, and Rhum, Gordon. The Effectiveness of College-Level Instruction In Freshman Composition. Interim Report. United States Office of Education Cooperative Research Project No. 2188. Cedar Falls, Iowa: State College of Iowa. December 1966.

Kincaid, Gerald L. "Some Factors Affecting Variations in the Quality of Students' Writing." Unpublished doctoral dissertation, Michigan State University, 1953.

Kitzhaber, Albert R. "New Perspectives on Teaching Composition," College English, 23 (March 1962), 440-444.

Kitzhaber, Albert R. Themes, Theories and Therapy: The Teaching of Writing in College. New York: McGraw Hill, 1963.

Lambert, Robert E. "The Use of Field Tryout Data." Paper presented bo the Annual Spring Conference, California Educational Research Associations, March 14, 1969.

McColly, William, and Remstad, Robert. "Composition Rating Scales for General Merit: An Experimental Evaluation," Journal of Educational Research, 59, No. 2 (October 1965), 55-56.

Miller, Joseph W. "An Analysis of Freshman Writing at the Beginning and End of a Year's Work in Composition." Unpublished doctoral dissertation, University of Minnesota, 1958.

Nealey, Stanley M. "Student-Instructor Agreement in Scoring an Essay Examination," Journal of Educational Research, 63, No. 3 (November 1969), 111-115.

Nisbet, J. D. "English Composition in Secondary School Selection," British Journal of Educational Psychology, 25 (February 1955), 51-54.

Nyberg, Verner Richard. "A Factor Analytic Study of Essay Gradings." Unpublished doctoral dissertation, University of California, Los Angeles, 1966.

Scannel, Dale P. and Haugh, Oscar M. Teaching Composition Skills With Weekly Multiple Choice Tests In Lieu Of Theme Writing. Final Report, United States Office of Education Project No. 6-8134. Washington, D. C. June, 1968.

Scriven, Michael. "The Methodology of Evaluation." In Perspectives of Curriculum Evaluation, edited by Tyler, Ralph W. et al. Chicago: Rand McNally, 1967.

Starch, Daniel and Elliott, Edward C. "Reliability of Grading High School Work in English," School Review, 20 (1912), 442-457.

Sutton, Joseph T. and Allen, Eliot. The Effects of Practice and Evaluation on Improvement in Written Composition. Cooperative Research Project No. 1993. Deland, Florida: Stetson University, 1964.

Torgerson, Warren S. and Green, Bert F. "The Factor Analysis of Subject Matter Experts," Journal of Educational Psychology, 43 (1953), 354-363.

Weingarten, Samuel and Kroeger, Frederick P. English In The Two-Year College. Champaign, Illinois: National Council of Teachers of English, 1965.

Zoellner, Robert. "Talk-Write: A Behavioral Pedagogy for Composition," College English, 30 (January 1969), 267-320.

Figure I: The Scoring Key

	DEGREE TO WHICH CRITERIA ARE MET			
	Meets all criteria (0 Point)	Average performance (1 Point)	Minimum acceptable performance (2 Points)	Unacceptable performance (3 Points)
SCORING DIRECTIONS	EACH PAPER SHOULD BE CHECKED (✓) FOR EACH ITEM 1-15			
CONTENT		0	1	2	3
. Treatment of subject	Unusually creative and perceptive	____	____	____	____
. Knowledge of subject	Obviously informed by experience or study	____	____	____	____
. Diction	Appropriate to content and style; exact word choice	____	____	____	____
ORGANIZATION					
. Thesis statement	Focus and purpose clearly identified	____	____	____	____
. Design or plan	Coherent and/or rhetorical objectives achieved	____	____	____	____
. Overall unity	All paragraphs contribute to development of thesis	____	____	____	____
. Development of individual paragraphs	Sufficiently developed with relevant, specific detail	____	____	____	____
. Sentence patterns	Coherent and varied patterns; good subordination/ coordination	____	____	____	____
. Transitions	All contribute to logical progress of ideas	____	____	____	____
. Logic	All ideas expressed rationally and without fallacy	____	____	____	____
MECHANICS					
. Spelling errors	None	____	____	____	____
. Syntax and word choice	Lucid, orderly arrangement; diction appropriate to context	____	____	____	____
. Sentence structure and punctuation and gross error penalties	Clear and grammatically correct	____	____	____	____
. Major mechanical errors	None	____	____	____	____
. Minor mechanical	None	____	____	____	____

GROSS ERROR PENALTIES: **Sentence Fragment = +8**